Docket No.: 388512010411
(PATENT)

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE
BEFORE THE BOARD OF PATENT APPEALS AND INTERFERENCES

In re Patent Application of:
Lawrence M. KAUVAR

Serial No.: 10/714,163 Confirmation No.: 2892

Filed: November 13, 2003 Art Unit: 1641

For: PROTEIN LOCALIZATION ASSAYS FOR TOXICITY AND ANTIDOTES THERETO Examiner: David J. Venci

BRIEF ON APPEAL

MS Appeal Brief - Patents
Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

Dear Sir:

A Notice of Appeal was filed in this case on 28 September 2005, thus setting a date for filing of the Brief of 28 November 2005. A petition for an extension of time of one (1) month until 28 December 2005 is enclosed along with the required fee. Claims 7-13 and 20-22 are subject to this Appeal.

1. Real Party in Interest

The real party in interest for this application is the assignee, Trellis Bioscience, Inc., having a current address in South San Francisco, California.

2. Related Appeals and Interferences

None of the appellant or his legal representative or the assignee is aware of any application, patent, appeal or interferences or judicial proceedings which would have a bearing on the decision in the pending appeal.

3. Status of Claims

There present application is a divisional of an application that is now U.S. Patent No. 6,673,554. It was filed with claims 1-20 and a Preliminary Amendment canceling claims 1-6 and 14-19. Claims 7-12 and 20 were amended in response to a First Office action and claims 21 and 22, depending from claim 7, were added. Thus, claims 7-13 and 20-22 are pending and are on appeal.

4. Status of Amendments

A proposed amendment was submitted in response to a final Office action in an attempt to accommodate outstanding rejections under 35 U.S.C. § 112, paragraph 2. That amendment would have inserted the limitations of claim 10 into independent claim 7 and substantially modified claim 22. It was proposed to cancel both claims 9 and 10. However, the amendment as proposed was not entered.

5. Summary of Claimed Subject Matter

Claims 7 and 20 are the independent claims in the present application.

Claim 7 is directed to a method to obtain a database of signal transduction protein localization profiles in response to toxic compounds (page 6, lines 3-5). The database is useful because the profiles in the database are surrogates for evaluating toxicity and thus provide a basis for evaluating the effectiveness of antidotes to toxins as well as using the toxin surrogate as a model for disease when similar profiles are obtained. The ability of a candidate antidote to alter the profile effected by a toxin in the database identifies the candidate as successful and the similarity of a toxin profile with that of a disease provides an identification method for treatments of the disease. (See page 4, lines 18-28 and page 6, lines 17-21.)

The steps require recording the intracellular localization pattern of at least one signal transduction protein and contacting each compound in a set of toxic compounds with a cell type that contains the signal transduction protein for which the intracellular localization pattern is to be obtained (page 6, lines 7-8). As the claim has been amended, the intracellular localization pattern must show the concurrent localization in at least three cellular locations selected from the group consisting of nuclear, perinuclear, diffuse cytoplasmic, cytoplasmic fibril-associated, and membrane-associated locations (page 10, lines 20-21 of the substitute specification filed 28 April 2004 and page 11, lines 1-2 of the specification as originally filed). At least three such locations must be concurrently analyzed such that profiles or footprints employing multiple data points are a feature of the invention. (See page 4, lines 19-22, for example.) Each intracellular location is recorded in computer readable and retrievable form as set forth in claim 7 as originally filed.

Obtaining these profiles that comprise at least three cellular locations is important because this provides a reliable surrogate to test compounds for their toxic effect by their ability to mimic the footprint or profile of a known toxin (page 15, lines 9-13).

The limitations of claims 8-12 are present in the claims as originally filed; the use of PKC isoenzymes as the signal transduction proteins (claim 8) is featured in the specification, for example, at page 6, beginning at line 27, and all of page 8. At least two or a multiplicity of signal transduction proteins (claims 9 and 10) may be assessed (page 6, lines 5-6). The use of a wide field microscope (claim 10) is a preferred method for assessment (page 9, beginning at line 25). Use of antibodies to track the signal transduction protein (claim 12) is a typical approach to labeling (page 10, lines 6-15). The limitations of claims 21 and 22 were originally in claim 7 as fled.

The other independent claim, claim 20, relates to an iterative method to identify a set of signal transduction proteins that are useful in the applications envisioned by the invention. (See page 6, lines 9-11.) The invention described in claim 20 is essentially that set forth in claim 20 as originally filed and only clarifying amendments were provided in the response to an initial Office action. Thus, the general procedure set forth in claim 20 was present in the application as originally filed.

6. Grounds of Rejection to be Reviewed on Appeal

Several aspects of claim wording were objected to under 35 U.S.C. § 112, paragraph 2. Appellant attempted to accommodate some of the Examiner's criticisms in the response to the final Office action, but entry of the Amendment was refused.

In claim 7, the objected-to phrases were: "each intracellular localization pattern" and "optionally as a function of time."

In claim 20, the terms objected to were "arbitrarily," "significant," "contacting," "each member," "adding new signal transduction proteins," "repeating the steps for which the second set of signaled transduction proteins was used," "range," "marketed" and "discarding those signal transduction proteins."

Claims 7, 9-10, 12-13 and 20-22 were rejected as assertedly anticipated under 35 U.S.C. § 102(e) by Dunlay and Taylor (U.S. Patent No. 5,989,835).

Claim 8 was rejected as assertedly obvious over Dunlay and Taylor in combination with Mochly-Rosen, *Science* (1995) 268:247.

Claim 11 was rejected as obvious over Dunlay and Taylor (*supra*) in view of Gerhardt (U.S. Patent No. 5,684,628).

Claim 20 was further rejected as obvious over Dunlay and Taylor (*supra*) in combination with Cook (U.S. Patent No. 6,546,378).

Copies of the cited documents are supplied with this Brief.

7. Argument

A. The claims as currently worded are sufficiently clear to meet the statutory requirement for definiteness.

Although appellant is willing to accommodate suggestions with regard to claim wording, attempts to do so in response to final rejection were not accepted. Nevertheless, even without the proposed amendments, the claims are clear and definite.

As to claim 7, "each" intracellular localization pattern is clear because although the claim requires only one such pattern in line 3, it admits of the possibility of many more. It appears clear as well that "recording" in line 8 of claim 7 is clearly a claim limitation and "optionally as a function of time" refers only to one possible method of recording.

As to claim 20, appellant does not agree that "arbitrarily" is a relative term requiring quantitative definition, but rather is indicative of random choice. Similarly, "significant" should be clear to any practitioner and in the context of the invention means that pattern changes are detectable in a reliable way above background. The word "contacting," in the context of the present invention, would most practicably be carried out by contacting the cells containing the signal transduction proteins; if way could be found to do so, it would not be detrimental to the invention to perform the contacting intracellularly. Either would be fine. Appellant is unable to understand the objection to "each member" of any set of signal transduction proteins. It does, of course, refer to a single signal transduction protein as each set is a collection of such individual members. Appellant also fails to understand the objections to the description of the iterative nature of the process in terms of "adding new signal transduction proteins" in line 12, since the original first set has been modified by discarding redundancies and a new set is created by adding more signal transduction proteins to be profiled. Similarly, "repeating the steps" is necessary because the process is iterative and new sets of signal transduction proteins are obtained after each round. As to "discarding" signal transduction proteins, these are simply no longer members of the set that will be members of the set identified according to the methods set forth in the claim. Finally, with respect to "the range" of "marketed" compounds, this refers to the variety in chemical space represented by compounds available in the real world.

While appellant believes that claim 20 as presented is sufficiently clear to meet legal standards, appellant agrees that the wording could be improved by an amendment which adds what appears to be an omitted step in front of the last paragraph of the claim, and an amendment after final has been submitted concomitantly herewith. A copy of the amendment is included as Exhibit A.

B. Claim 20 is clearly not anticipated by Dunlay and Taylor.

No rationale has been provided for the rejection of claim 20 as anticipated by Dunlay and Taylor, and appellant is unable to find any disclosure therein that describes or suggests the method of claim 20. There appears to be no disclosure in Dunlay and Taylor of any iterative process to select a set of signal transduction proteins which will provide at least five principal components with respect to the range of compounds marketed as small organic molecules, or any iterative process at all.

C. Claim 7 is not anticipated by Dunlay and Taylor nor is the method of claim 7 suggested by this document.

Claim 7 is directed specifically to testing sets of toxic compounds to obtain a database of cellular localization patterns for a variety of toxins. Each localization pattern in the database must assess at least three cellular locations selected from nuclear, perinuclear, diffuse cytoplasmic, cytoplasmic fibril associated, and membrane associated locations.

Dunlay and Taylor is primarily concerned with providing a spatial arrangement of a multiplicity of cell samples so that high throughput screening can be conducted. Essentially, it provides microtiter plates so that multiplicities of cell samples can be assessed efficiently using a

specific system of assessing the effects of any biological stimulus. There is nothing in this document that is concerned with providing translocation profiles for multiplicities of compounds consisting of toxins. Rather, the disclosure of Dunlay and Taylor is directed to a generic assay system for intracellular translocation, the subject of which may or may not be signal transduction proteins (signal transduction proteins are included but not the focus of the Dunlay and Taylor system). In one instance where a signal transduction protein is selected in a hypothetical Example 1 in column 9, only two cellular locations are evaluated, not three as required by the claim. Similarly, in discussing the possibility of viewing the translocation of NF-κβ in column 7, beginning at line 55, only two locations are evaluated.

Thus, there are two basic differences, at least, between the claimed invention of claim 7 and the cited document - claim 7 requires providing a set of toxic compounds specifically and evaluating the localization patterns in at least three specified cellular locations.

No citation should be required to support the proposition that, in order for anticipation to be found, each and every claim limitation must be met. Appellant has pointed out two distinctions between the cited document and claim 7, thus the rejection for anticipation is improper.

Neither does Dunlay and Taylor suggest the invention as claimed. The invention is directed to a method to obtain a database of signal transduction protein localization profiles specifically in response to toxic compounds and Dunlay nowhere suggests such a method; Dunlay and Taylor is concerned with the mechanics of localizing cellular components in general and recording the results in a high throughput format, and is not concerned with the method of the invention directed to studying toxicity by employing localization profiles of signal transduction proteins in at least three locations.

As claim 7 is free of anticipation by Dunlay and Taylor, so too are its dependent claims 9-10, 12-13 and 20-22.

D. Claims 8 and 11 are patentable over the cited combinations.

Claim 8 requires that at least one of the signal transduction proteins be a protein kinase C (PKC) isoenzyme. Appellant agrees that PKC is a useful signal transduction protein as taught by Mochly-Rosen; however, as the method in which it is employed is not suggested by the art as argued above, the combination of Dunlay and Taylor with Mochly-Rosen does not suggest the invention of claim 8.

Claim 11 requires the use of a wide field microscope. Appellant agrees that Gerhardt, among other possible citations describing such instruments, identifies the wide field microscope as a useful tool in evaluating intracellular locations. The method of claim 7, however, is not suggested by the primary document, so the inclusion of the wide field microscope in that method cannot be used to defeat patentability.

E. Claim 20 is not rendered obvious even if Dunlay and Taylor is combined with Cook.

As a preliminary matter, it is difficult to see why the artisan practicing in the field of evaluating the effects of toxic compounds on intracellular localization patterns of signal transduction proteins would be motivated to consult the Cook publication, which appears to be in the field of signal interpretation in general, a specialty derived from an entirely different discipline. It appears on its face that Cook is not analogous art, which is not properly cited in the context of the present invention, or indeed in the context of Dunlay and Taylor. Cook is not in the field of the invention, nor is it pertinent to the problem to be solved. *In re Deminski*, 796 F2d 436,

230 USPQ 313 (Fed. Cir. 1986). However, whether analogous or not, there appears to be no motivation to combine the two documents.

The three appropriate criteria for combination are set forth in *In re Rouffet*, 47 USPQ2d 1453 (Fed. Cir. 1998) – there must be a suggestion in the documents themselves (not present here), the problem to be solved must be similar (not present here) or one of the documents must be particularly high profile (also not the case here). Notably, the Examiner has not provided any rationale whatsoever for combining these documents, which, as made clear in *Rouffet*, is an absolute requirement for maintaining a rejection based on a combination of documents. In sum, combining Dunlay and Taylor with Cook is unjustified on the record.

But even if the combination is made, the invention as claimed is not suggested. The rejection as it stands assumes that Dunlay and Taylor suggests all elements of claim 20 except the element of principal component analysis. That is not the case, as was discussed in detail above - Dunlay and Taylor fail to suggest anything remotely resembling claim 20, and no description of how claim 20 is suggested by Dunlay and Taylor has ever been provided. Claim 20 is an iterative method of identification to obtain a set of signal transduction proteins with profiles that provide the required characteristic of five principal components with respect to the range of compounds marketed as small organic molecules. And of course, nothing in Cook does anything other than mention principal component analysis in a list of possible analysis of "composition matrices." There is no particular nexus between this mention of "principal component" analyses, and "processing data from pharmaceutical/drug effect studies" which also appears on a long list of possible applications of a signal interpretation method and apparatus as described in Cook. The second reference made by the Examiner to a purported suggestion to use "principal components"

to test and evaluate drugs at column 16, line 30, which would at least be proximal in space to the mention of "principal components" appears to be nonexistent. In short, Cook cannot be said specifically to suggest the application of principal component analysis to data from pharmaceutical/drug effect studies much less to suggest the use of principal component analysis to be applied to either the method of Dunlay and Taylor or the method of claim 20 which methods themselves are quite distinct from each other.

Further, claim 20 does not even require a step of principal component analysis, other than to assure that the set of signal transduction proteins identified by the claimed method meets the criterion that it has this desired property.

The combination of Dunlay and Taylor with Cook is inappropriate and even if made falls far short of suggesting the method of claim 20.

F. Conclusion

Appellant respectfully requests the foregoing rejections be reversed and claims 7-13 and 20-22 be indicated allowable.

8. Claims Appendix

An Appendix containing a copy of the claims as currently pending is attached.

9. Evidence Appendix

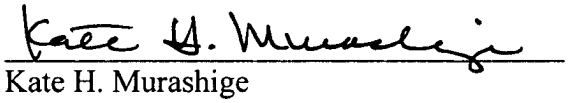
An Appendix containing copies of Exhibits A and B submitted with the response to the first Office action is attached.

10. Related Proceedings Appendix

No related proceedings are referenced in 2. above, therefore no Appendix is included.

The Assistant Commissioner is hereby authorized to charge any additional fees under 37 C.F.R. § 1.17 that may be required by this Brief, or to credit any overpayment, to **Deposit Account No. 03-1952**.

Respectfully submitted,

Dated: December 28, 2005 By: 

Kate H. Murashige
Registration No. 29,959

Morrison & Foerster LLP
12531 High Bluff Drive
Suite 100
San Diego, California 92130
Telephone: (858) 720-5112
Facsimile: (858) 720-5125

CLAIMS APPENDIX

1-6. (canceled)

7. (previously presented): A method to obtain a database of signal transduction protein localization profiles in response to toxic compounds, which method comprises

recording the intracellular localization pattern of at least one signal transduction protein in a cell type,

providing a set of toxic compounds,

contacting each compound of said set of toxic compounds with said cell type,

recording the intracellular localization pattern of at least one of said signal transduction proteins in said cell type in the presence of each compound in said set of toxic compounds, optionally as a function of time,

wherein each intracellular localization pattern is constructed by concurrently determining the presence, absence or amount of said signal transduction protein in at least three cellular locations selected from the group consisting of nuclear, perinuclear, diffuse cytoplasmic, cytoplasmic fibril-associated, and membrane-associated locations;

wherein each intracellular localization pattern is recorded in computer-readable and retrievable form.

8. (previously presented): The method of claim 7 wherein at least one of said signal transduction proteins is a protein kinase C (PKC) isoenzyme.

9. (previously presented): The method of claim 7 wherein the intracellular localization patterns of at least two signal transduction proteins are determined.

10. (previously presented): The method of claim 9 wherein the intracellular localization patterns of a multiplicity of signal transduction proteins are determined.

11. (previously presented): The method of claim 7 wherein each of said intracellular localization patterns is observed using a wide-field microscope.

12. (previously presented): The method of claim 7 wherein each of said intracellular localization patterns is observed by labeling the proteins with specific antibodies.

13. (original): A computer-readable database prepared by the method of claim 7.

14-19. (canceled)

20. (previously presented): A method to identify a set of signal transduction proteins whose intracellular localization pattern changes significantly in response to toxic compounds, which method comprises

arbitrarily identifying a first set of signal transduction proteins;

providing a set of toxic compounds;

contacting each member of said first set of signal transduction proteins with each one of the toxic compounds;

determining the changes in intracellular localization pattern of each of the signal transduction proteins of said first set in response to each of the toxic compounds;

discarding those signal transduction proteins from said first set whose changes in intracellular localization pattern are redundant;

adding new signal transduction proteins to provide a second set of signal transduction proteins;

contacting each member of said second set of signal transduction proteins with each of the toxic compounds;

determining the changes in the intracellular localization pattern of each of the signal transduction proteins of said second set in response to each of the toxic compounds;

discarding those signal transduction proteins from said second set whose changes in intracellular localization patterns are redundant; and

repeating the steps for which the second set of signal transduction proteins was used until a final set of proteins is obtained which provides at least five principal components with respect to the range of compounds marketed as small organic molecules.

21. (previously presented): The method of claim 7, which further includes the step of recording the intracellular localization pattern of said signal transduction protein in said cell type in the presence of each compound of said set of toxic compounds as a function of time.

22. (previously presented): The method of claim 7, which further includes the step of recording the intracellular localization pattern of said signal transduction protein in said cell type, then contacting each compound of said set of toxic compounds with a second cell type, and recording the intracellular localization pattern of said first signal transduction protein in said second cell type in the presence of each compound of said set of toxic compounds.

EVIDENCE APPENDIX

This appendix contains, for the convenience of the Office, the following evidentiary material already of record:

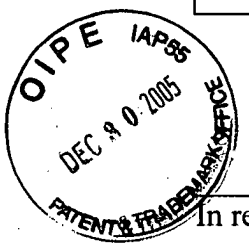
1. An excerpt from *Chemical and Engineering News*, February 12, 1996, describing combinatorial chemistry; and

2. Gieni, *et al., J. Inf. Chem. Comput Sci.* (1999) 39:1076-1080 which is an article showing principal components analysis.

I hereby certify that this correspondence is being deposited with the U.S. Postal Service with sufficient postage as First Class Mail, in an envelope addressed to: MS AF, Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450, on the date shown below.

Dated: December 28, 2005 Signature: /Marian L. Christopher/
(Marian L. Christopher)

Docket No.: 388512010411
(PATENT)



IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re Patent Application of:
Lawrence M. KAUVAR

Serial No.: 10/714,163 Confirmation No.: 2892

Filed: November 13, 2003 Art Unit: 1641

For: PROTEIN LOCALIZATION ASSAYS FOR TOXICITY AND ANTIDOTES THERETO Examiner: David J. Venci

AMENDMENT UNDER 37 C.F.R. § 1.116

MS AF
Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

Dear Sir:

The following amendment is proposed upon reflection with regard to the wording of claim 20. It is believed that a step, which would be apparent to the reader, has been inadvertently omitted and that the claim would be clarified by correcting this omission. Reconsideration is respectfully requested.

CLAIM AMENDMENTS

1-6. (canceled)

7. (previously presented): A method to obtain a database of signal transduction protein localization profiles in response to toxic compounds, which method comprises

recording the intracellular localization pattern of at least one signal transduction protein in a cell type,

providing a set of toxic compounds,

contacting each compound of said set of toxic compounds with said cell type,

recording the intracellular localization pattern of at least one of said signal transduction proteins in said cell type in the presence of each compound in said set of toxic compounds, optionally as a function of time,

wherein each intracellular localization pattern is constructed by concurrently determining the presence, absence or amount of said signal transduction protein in at least three cellular locations selected from the group consisting of nuclear, perinuclear, diffuse cytoplasmic, cytoplasmic fibril-associated, and membrane-associated locations;

wherein each intracellular localization pattern is recorded in computer-readable and retrievable form.

8. (previously presented): The method of claim 7 wherein at least one of said signal transduction proteins is a protein kinase C (PKC) isoenzyme.

9. (previously presented): The method of claim 7 wherein the intracellular localization patterns of at least two signal transduction proteins are determined.

10. (previously presented): The method of claim 9 wherein the intracellular localization patterns of a multiplicity of signal transduction proteins are determined.

11. (previously presented): The method of claim 7 wherein each of said intracellular localization patterns is observed using a wide-field microscope.

12. (previously presented): The method of claim 7 wherein each of said intracellular localization patterns is observed by labeling the proteins with specific antibodies.

13. (original): A computer-readable database prepared by the method of claim 7.

14-19. (canceled)

20. (currently amended): A method to identify a set of signal transduction proteins whose intracellular localization pattern changes significantly in response to toxic compounds, which method comprises

arbitrarily identifying a first set of signal transduction proteins;

providing a set of toxic compounds;

contacting each member of said first set of signal transduction proteins with each one of the toxic compounds;

determining the changes in intracellular localization pattern of each of the signal transduction proteins of said first set in response to each of the toxic compounds;

discarding those signal transduction proteins from said first set whose changes in intracellular localization pattern are redundant;

adding new signal transduction proteins to provide a second set of signal transduction proteins;

contacting each member of said second set of signal transduction proteins with each of the toxic compounds;

determining the changes in the intracellular localization pattern of each of the signal transduction proteins of said second set in response to each of the toxic compounds;

discarding those signal transduction proteins from said second set whose changes in intracellular localization patterns are redundant;

adding new signal transduction proteins to provide a third set of signal transduction proteins; and

repeating the steps for which the second set of signal transduction proteins was used until a final set of proteins is obtained which provides at least five principal components with respect to the range of compounds marketed as small organic molecules.

21. (previously presented): The method of claim 7, which further includes the step of recording the intracellular localization pattern of said signal transduction protein in said cell type in the presence of each compound of said set of toxic compounds as a function of time.

22. (previously presented): The method of claim 7, which further includes the step of recording the intracellular localization pattern of said signal transduction protein in said cell type, then contacting each compound of said set of toxic compounds with a second cell type, and recording the intracellular localization pattern of said first signal transduction protein in said second cell type in the presence of each compound of said set of toxic compounds.

REMARKS

It appears helpful to make explicit the implied step that additional proteins would be added to the set before repeating the iterative steps. The proposed amendment is implicit in the claims as currently proposed and thus entry of the amendment is believed proper.

In the unlikely event that the transmittal letter is separated from this document and the Patent Office determines that an extension and/or other relief is required, applicants petition for any required relief including extensions of time and authorize the Assistant Commissioner to charge the cost of such petitions and/or other fees due in connection with the filing of this document to **Deposit Account No. 03-1952** referencing docket No. 388512010411.

Respectfully submitted,

Dated: December 28, 2005 By: Kate H. Murashige

Kate H. Murashige
Registration No. 29,959
MORRISON & FOERSTER LLP
12531 High Bluff Drive
Suite 100
San Diego, California 92130-2040
Telephone: (858) 720-5112
Facsimile: (858) 720-5125

Chemical & Engineering News,
February 12, 1996

SPECIAL REPORT

Combinatorial chemists focus on small molecules, molecular recognition, and automation

Stu Borman,

C&EN Washington

Drug candidates traditionally have been synthesized one at a time, a time-consuming and labor-intensive process. But many researchers in academia, government, biotechnology firms, and drug companies increasingly are turning to combinatorial chemistry - a strategy for creating new drugs that, it is hoped, will speed the drug discovery process significantly.

The idea of combinatorial chemistry is to make a large number of chemical variants all at one time; to test them for bioactivity, binding with a target, or other desired properties; and then to isolate and identify the most promising compounds for further development.

The success of combinatorial chemistry is still uncertain. No drugs discovered combinatorially have been approved for marketing, although several are currently in development. But many researchers believe the technique will prove to be an efficient and cost-effective tool for identifying new medicines.

In combinatorial chemistry experiments, chemical libraries (large collections of compounds of varied structure) are produced by sequentially linking different molecular building blocks, or by adding substituent "decorations" to a core structure such as a polycyclic compound. Libraries may consist of molecules free in solution, linked to solid particles or beads, or even arrayed on surfaces of modified microorganisms.

Combinatorial chemistry initially focused on the synthesis of very large libraries of biological oligomers such as peptides and oligonucleotides. But drug developers generally prefer to focus on small organic molecules with molecular weights of about 500 daltons or less - the class of compounds from which most successful drugs have traditionally emerged. So combinatorial chemistry researchers are concentrating on small organic compounds as well.

Drug discovery is the primary goal of most combinatorial chemistry research, but combinatorial methods also have potential applications for development of advanced materials and catalysts.

One of the challenges of combinatorial chemistry is the difficulty of identifying "hits" (active compounds) present at vanishingly low concentrations in complex combinatorial libraries. To address this problem, ingenious encoding schemes have been developed. Two groups have independently

developed the latest concept in this field - radiofrequency encoding, in which information about library compounds is stored on microchips.

Instrumentation systems to help speed combinatorial chemistry experiments have been developed in-house at a number of biotechnology and pharmaceutical companies. And several combinatorial automation systems are available commercially or undergoing intensive development.

Combinatorial chemistry has come a long way in just a few years, but further advances are needed and new applications are anticipated. Directions in which the field is headed range from combining combinatorial chemistry with computational drug-design strategies to the use of combinatorial molecular recognition for studies of protein function.

Creating libraries

Combinatorial libraries are created in the laboratory by one of two methods - split synthesis or parallel synthesis. In split synthesis, compounds are assembled on the surfaces of microparticles or beads.In each step, beads from previous steps are partitioned into several groups and a new building block is added. The different groups of beads are then recombined and separated once again to form new groups. The next building block is added, and the process continues until the desired combinatorial library has been assembled.

Before split synthesis was developed, explains chemistry professor Kim D. Janda of Scripps Research Institute, La Jolla, Calif., "people created diversity using mixtures of compounds. In a coupling step, you would add, let's say, reagents A, B, and C in one pot, and A, B, and C would all compete to become integrated at the same site. But in doing that you can have problems with kinetics. One reaction may be faster than another and you may not get equal distribution of the three components."

Split synthesis "got away from all that," says Janda. "You could create diversity using separate reactions, so the components would have an equal chance to add in to a site, and then by mixing compounds together again you got the diversity you needed."

Libraries resulting from split synthesis are characterized by the phrase "one bead, one compound." Each bead in the library holds multiple copies of a single library member. Split synthesis greatly simplifies the isolation and identification of active agents because beads (and implicitly individual library members) are large enough to be observed visually and separated mechanically.

Combinatorial libraries can also be made by parallel synthesis, in which different compounds are synthesized in separate vessels (without remixing), often in an automated fashion. Unlike split synthesis, which requires a solid support, parallel synthesis can be done either on a solid support or in solution.

A commonly used format for parallel synthesis is the 96-well microtiter plate. Robotics instrumentation can be used to add different reagents to separate wells of a microtiter plate in a predefined manner to produce combinatorial libraries. Hits from the library can then be identified by well location.

Split synthesis is used to produce small quantities of a relatively large number of compounds, whereas parallel synthesis yields larger quantities of a relatively small number of compounds. And split synthesis requires that assays be performed on pools of compounds, whereas assays on individual compounds can be run on libraries created by parallel synthesis. While slower, testing individual compounds is sometimes advantageous because serious interferences and complications can arise when multiple compounds are tested simultaneously.

A special case of parallel synthesis is spatially addressable synthesis, pioneered by researchers at Affymax Research Institute, Palo Alto, Calif. In this technique, libraries are synthesized in arrays on microchips, and all the compounds on a chip are assayed simultaneously for binding or activity. Hits can then be identified by the piece of real estate they occupy on the chip. Using a chip-making technique called photolithography, Affymax researchers have generated arrays of more than 65,000 compounds on chips about 1 sq cm in area.

Bioactive combinatorial compounds synthesized by split synthesis can also be identified by deconvolution, a technique in which each variable position in a compound library is tested to find the building block that makes the strongest contribution to activity at that site.

Solid-phase and solution-phase combinatorial synthesis each have their advantages and disadvantages. Solid-phase synthesis permits use of excesses of reagents to drive reactions to completion, since excess reagents can be washed away from beads very easily afterward. However, solution-phase synthesis is more versatile because many organic solution-based reactions have not been adapted for solid-phase work.

Janda and coworkers at Scripps recently developed a liquid-phase synthesis procedure that combines some of the advantages of solution-phase and solid-phase synthesis [*Proc. Natl. Acad. Sci. USA,* **92,** 6419 (1995)]. The procedure involves use of polyethylene glycol monomethyl ether in place of solid-phase beads as a foundation for combinatorial assembly. The polymer is soluble in a variety of aqueous and organic solvents, making it possible to use solution-phase combinatorial synthesis. But the polymer can be precipitated out of solution by crystallization at each stage of the combinatorial process to facilitate purifications.

Small-molecule libraries

Combinatorial chemistry began with the synthesis of large libraries of biopolymers such as peptides and oligonucleotides. In some cases, these were created on surfaces of genetically modified microorganisms, such as bacteriophage particles, by inserting combinatorial DNA oligomers into genes that encode cell-surface proteins.

However, peptides and oligonucleotides are problematic for drug development because their oral bioavailability is poor and they are degraded rapidly by enzymes. Hence, the focus of combinatorial research has shifted in recent years to libraries of nonpolymeric small molecules having molecular weights of about 500 daltons or less.

In a pioneering study, chemistry professor Jonathan A. Ellman and coworkers at the University of California, Berkeley, synthesized the first such library by creating variants of benzodiazepines, a class of compounds that has been a fertile source of successful drugs [*J. Am. Chem. Soc.,* **114,** 10997 (1992)]. Since then, researchers have found ways to synthesize combinatorial libraries based on many other classes of small organic compounds.

A recent example is work by Mark A. Gallop, director of combinatorial chemistry, and coworkers at Affymax. They used a cycloaddition reaction to prepare a small-molecule combinatorial library of about 500 mercaptoacyl prolines [*J. Am. Chem. Soc.,* **117,** 7029 (1995)]. By screening this library, they identified an unusually potent inhibitor of angiotensin-converting enzyme (ACE). ACE inhibitors are used as treatments for hypertension and heart disease.

And the group of Stephen W. Kaldor, head of combinatorial chemistry research at Eli Lilly & Co.,

Indianapolis, in collaboration with scientists in Lilly's central nervous system (CNS) group, has used combinatorial chemistry to identify an orally active CNS agent by combinatorial optimization of an existing lead. The low molecular weight nonoligomeric drug candidate entered clinical trials in November. "This is one of the first small-molecule combinatorial compounds to go into humans," says Kaldor.

A major challenge of small-molecule combinatorial chemistry has been to adapt conventional solution-phase organic reactions to reactions on solid-phase particles. Ellman says one of his group's efforts "has been to expand the kind of chemistry that can be performed on solid supports in a simultaneous synthesis format - in particular, carrying out different types of carbon-carbon bond-forming reactions. For example, we've developed general enolate alkylation conditions where side reactions that can be a problem in solution don't occur."

Paralleling the increasing use of small-molecule libraries is a trend toward assaying libraries having smaller numbers of components. "People seem to be much more comfortable working with smaller mixtures - probably a hundred components or less in a mixture, rather than the mixtures of 10^5 and 10^6 compounds per pool that we saw in the early experiments," says Ronald N. Zuckermann, associate director of bioorganic chemistry at Chiron Corp., Emeryville, Calif. "The lower the number of compounds, the more confidence you can have in the biological data" because artifacts arise more readily in the screening of large pools of compounds.

Ellman agrees that "people have gotten away from screening really large mixtures of compounds. They either want to screen them individually or in smaller pools of under 100 compounds. It's easier to extract out binding data in that format."

Oligomers and materials

Carbohydrates have lagged behind other types of compounds in combinatorial library development because of the complexity of oligosaccharide chemistry, but carbohydrate libraries are now beginning to appear. For example, Ole Hindsgaul and coworkers at the department of chemistry of the University of Alberta, Edmonton, in collaboration with researchers at the University of Georgia, Athens, and Ciba Central Research Laboratories, Basel, Switzerland, have developed a "random glycosylation" strategy for making oligosaccharide libraries in solution [*Angew. Chem. Int. Ed. Engl.*, **34,** 2720 (1995)]. They produced a library of all 18 possible fucosylated trisaccharides from disaccharide precursors.

And at a recent meeting, chemistry professor Daniel E. Kahne of Princeton University reported construction of the first solid-phase carbohydrate library, using chemistry for solid-phase synthesis of oligosaccharides developed earlier by his group. This technique has been licensed to Transcell Technologies, Monmouth Junction, N.J. In preliminary work, compounds isolated from one carbohydrate library have been shown to bind a carbohydrate-binding protein with greater affinity than the protein's natural ligand. "Carbohydrates play a central role in some very important biological processes, so having access to libraries of these compounds is critical," says Kahne.

Another type of oligomer being pursued combinatorially is peptoids, peptide analogs that are not recognized by peptide-cleaving enzymes. Chiron researchers recently discovered a candidate urokinase receptor antagonist from a peptoid library, and the compound is currently in preclinical studies as a potential anticancer agent.

"One of the primary advantages of peptoids is their synthetic accessibility," says Zuckermann. "They are efficiently synthesized by the submonomer method, which uses primary amines and bromoacetic acid as

starting materials - both very cheap, and there are literally thousands of amines readily available. The combination of this chemistry with robotic synthesis has led to a truly high throughput synthesis facility."

Chiron's identification of nanomolar peptoids that bind to transmembrane receptors [*J. Med. Chem.*, **37**, 2678 (1994)] "was the first example of the discovery of potent ligands to pharmaceutically relevant receptors from a combinatorial library of nonpeptides or nonnucleic acids - that is, synthetic compounds," Zuckermann adds. "I believe that this work helped inspire others to continue to move away from peptides and further toward small molecules."

Combinatorial chemistry can also be extended entirely beyond the realm of organic chemistry. For example, physicist Xiao-Dong Xiang of Lawrence Berkeley National Laboratory, chemistry professor Peter G. Schultz of UC Berkeley, and coworkers recently devised a combinatorial strategy for finding advanced materials with novel chemical or physical properties - extending "the combinatorial approach from biological and organic molecules to the remainder of the periodic table," as they put it [*Science*, **268**, 1738 (1995)].

Xiang, Schultz, and coworkers used thin-film deposition and physical masking techniques to synthesize libraries of solid-state materials. The properties of the resulting materials were then evaluated to identify promising candidates for further development.

Encoding

In spatially addressable combinatorial synthesis, active compounds can be identified by location. But in other forms of combinatorial chemistry, identifying hits is not so easy because there's often too little of each compound present for characterization with traditional analytical chemistry techniques.

Hence, many researchers now use some form of tagging or encoding to label compounds in large combinatorial libraries. The first such encoding scheme was proposed in 1992 by Scripps President Richard A. Lerner and molecular biologist Sydney Brenner at the institute. They suggested that a combinatorial library could be encoded with oligonucleotides synthesized in parallel with library compounds and linked to each one. Amplification or decoding of the attached oligonucleotide would serve to identify the small molecule bound to each bead.

This idea was independently arrived at and reduced to practice by scientists at Affymax. Later on, researchers at Chiron and at Selectide, Tucson, Ariz., developed similar techniques in which peptides instead of oligonucleotides were used as the sequenceable encoding oligomers.

In 1993, chemistry professor W. Clark Still and coworkers at Columbia University developed a second major type of encoding scheme, in which chromatographically resolvable organic tags were used as encoding elements for bead-based combinatorial libraries. Still devised the technique in response to concerns about the tendency of DNA and peptide tags to break down under the often very rough conditions of organic synthesis.

In Still's technique, inert halogenated aromatic compounds are used to encode the chemical reaction history experienced by each bead. These tags are identified by capillary gas chromatography to reveal the identity of active compounds in the library. Kahne, who used this type of encoding to construct his combinatorial carbohydrate library, says the method "is as good as it gets for identifying hits - a very simple solution to a very important problem."

The most recent development in encoding technology involves the use of radiofrequency tags. Chemistry professor K. C. Nicolaou at Scripps and the University of California, San Diego, together with senior chemist Xiao-Yi Xiao, President and Chief Executive Officer Michael P. Nova, and their coworkers at IRORI Quantum Microchemistry, La Jolla, Calif., developed a technique in which memory devices are associated or coated directly with derivatized polymer during combinatorial synthesis [*Angew. Chem. Int. Ed. Engl.*, **34,** 2289 (1995)]. The chips encode relevant information about the synthetic pathway - including not only reagents used, but also reaction conditions such as temperature and pH. The device can then "report" this information to a receiver via radiofrequency transmission.

"We're putting a manual system to do this type of radiofrequency combinatorial chemistry out on the market in March," says Nova. The system will include radiofrequency memory devices in MicroKans, tiny spherical capsules with porous walls that also enclose polymer beads for combinatorial synthesis.

A related technique was developed independently by synthetic chemist Edmund J. Moran and coworkers at Ontogen Corp., Carlsbad, Calif., and the University of California, Los Angeles [*J. Am. Chem. Soc.*, **117,** 10787 (1995)]. This approach differs from the Scripps technique in that reaction data from each stage of combinatorial synthesis are stored in a computer database, rather than being retained in the chip itself. An identification number stored in the memory of each chip is a pointer to reaction information in the database. Moran and coworkers have applied the strategy successfully to the discovery of novel inhibitors of a protein tyrosine phosphatase.

Molecular recognition

A combinatorial chemistry application that has become increasingly active in the past year or so, and that promises to grow even more rapidly in the future, is combinatorial molecular recognition - the use of combinatorial techniques to study binding between biological or synthetic receptors and their ligands. Researchers in the combinatorial molecular recognition community "want to be able to make small molecules that do the kinds of things that antibodies do - tightly and selectively bind important molecules or transition states," explains Columbia's Still.

"We made libraries of substrates just to measure the binding properties of compounds synthesized as enantioselective receptors," says Still, "and the receptors did indeed have significant sequence-selective binding properties that had never been observed before." The results of such experiments suggest, says Still, "that virtually anything people can do with antibodies ought to be doable with small molecules, and that it may not be that hard to identify small molecules that are as selective as antibodies for binding substrates."

Chemistry professor Stuart L. Schreiber and coworkers at Harvard University are also using combinatorial molecular recognition - in this case in conjunction with nuclear magnetic resonance spectroscopy (NMR) - to study protein receptors. They have focused initially on the SH3 domain, a frequently occurring structural feature in proteins (such as tyrosine kinases) involved in signal transduction.

The researchers identified peptide ligands that bound SH3 in two binding pockets that make up part of the SH3 binding site. The SH3 binding site also includes a third binding pocket that is highly variable in structure and is therefore referred to as a "specificity pocket." A combinatorial strategy led to the discovery of two classes of peptide ligands that bind to the three pockets in opposite orientations, as determined by NMR analysis of the SH3-ligand complexes. Last month, Schreiber and coworkers reported also having identified nonpeptide elements that bind to the specificity pocket [*J. Am. Chem. Soc.*, **118,** 287 (1996)].

Chemistry professor Fredric M. Menger and coworkers at Emory University, Atlanta, are also using a form of combinatorial molecular recognition - in this case to identify industrial catalysts [*J. Org. Chem.*, **60,** 6666 (1995)].

"Libraries have in the past been screened for noncovalent binding," says Menger. "But this is the first, or one of the first, cases where purely organic libraries have been investigated for catalytic activity. We make hundreds or thousands of compounds very quickly and then test their catalytic power. The potential catalysts are polymers that have multiple functional groups in different proportions and different sequences, plus a metal ion."

In screening for catalytic activity, "we selected the hydrolysis of a phosphate ester," says Menger, "but one could choose any reaction of interest. Once a polymer with activity is found, we begin tinkering with the proportions to fine-tune it until it gets faster and faster."

Using this approach, Menger and coworkers have identified polymers that accelerate phosphate hydrolysis by a factor of 10^4 or more. According to Menger, "The potential exists for even greater acceleration ... since only a small portion of the vast number of possible combinations has as yet been tested."

In future work, the researchers hope to make chiral polymers that can reduce functional groups enantioselectively. "I would not be surprised if in 10 years most new catalysts are developed combinatorially," says Menger. "Industry is moving more and more toward aqueous systems to avoid organic solvents. If one could devise catalysts for organic reactions in water, that would be a useful practical development."

Automation

Planning and performing combinatorial experiments in the laboratory is a complex and potentially tedious process. Hence, "A future trend is going to be greater availability of automation devices," says Chiron's Zuckermann. "A lot of solutions are being developed for automating combinatorial split synthesis or multiple parallel synthesis."

For example, Chiron has developed proprietary robotic combinatorial synthesizers. "We now have third-generation units working in our labs that feature all-glass reaction vessels, heating to 120°C, and flexible software, [allowing] automation of most organic reactions," says Zuckermann.

Ontogen has developed OntoBLOCK, an in-house combinatorial chemistry automation system that can produce 1,000 to 2,000 small organic molecules per day by parallel array synthesis. The system includes reaction blocks containing 96 reaction vessels, from which compounds can be transferred directly to standard 96-well microtiter plates for high-throughput screening.

Bohdan Automation Inc., Mundelein, Ill., markets a combinatorial chemistry reaction block that accommodates a wide variety of organic solvents and handles both solid-phase and solution-phase chemistry. Advanced ChemTech, Louisville, markets instrumentation for combinatorial peptide and organic synthesis. And Tecan U.S. Inc., Research Triangle Park, N.C., offers an organic chemical synthesizer called CombiTec that includes a robotic sample processor and reaction blocks of eight to 56 chambers.

Robotics maker Zymark Corp., Hopkinton, Mass., has put together several different automation systems that enable their clients to do solution-phase combinatorial synthesis and solid-phase peptide and

peptoid synthesis. The reactions can generally be performed under inert gas at a variety of temperatures.

According to Brian Lightbody, general manager of drug discovery business development at Zymark: "The process of generating combinatorial compounds involves several steps in addition to the actual reaction - [including] initial formulation of the reactants, labeling, pooling and splitting, cleavage, liquid-liquid extraction, solid-phase extraction, and evaporation. These steps require extensive manual labor. ... An automated robotic approach can often be implemented to fulfill these requirements, dramatically reducing the manual labor and eliminating the sources of human error."

A combinatorial chemistry system still in the prototype stage is the Nautilus, a synthetic chemistry workstation being developed by Argonaut Technologies Inc., San Carlos, Calif. The instrument handles a wide range of reagents, with capabilities for temperature control and use of inert atmospheres.

"The Nautilus allows you to do pretty much what you're able to do on the bench except in an automated fashion," says Argonaut President and CEO Joel F. Martin. "The system is completely enclosed and encapsulated, with a pressurized fluid delivery system and no exposure to the atmosphere whatsoever. It's a closed system, and all wetted surfaces within the instrument are glass or [polytetrafluoroethylene, such as DuPont's] Teflon."

Procedures that have been demonstrated on the Nautilus include a Suzuki coupling (a carbon-carbon bond-forming reaction at elevated temperature using an air-sensitive palladium catalyst), a butyllithium reaction, enolate reactions of the type developed by Ellman and coworkers, and synthesis of a solid-phase druglike molecule. "We chose tough organic reactions that no one would ever have conceived of doing in an automated synthesizer in the past," says Martin. The Nautilus is scheduled to be released commercially in August.

CombiChem Inc., San Diego, is developing commercial instrumentation for combinatorial chemistry that is likely to be competitive with the Nautilus. "Every company now is looking at ways of automating synthesis, purification, and analysis," says CombiChem Chief Operating Officer Peter L. Myers. "There's a major revolution going on. It's probably not obvious to a lot of people, and the academics may think we're overemphasizing it. But I know for a fact that every company now is looking to automate combinatorial chemistry because chemistry's become the rate-determining step."

As to whether conventional robotics instrumentation can be used effectively for combinatorial chemistry synthesis, "This immediately gets one into a debate," replies Myers. "When you're doing chemical reactions to make small molecules, many of the reactions are sensitive to conditions such as the presence of water vapor or oxygen, so inert atmospheres such as argon and nitrogen are often needed. The only successful way of blanketing a reaction is to have a closed system - one that is sealed. And if you seal it, then of course you can't use a robot very easily."

Myers adds, "This is why we, and also Argonaut, have gone to nonrobot systems - closed systems that work on valves and plumbing. ... That essentially means individual reaction vessels presealed with a solid support or chemicals inside, delivery by valving, and some way to agitate or stir the contents. Then you let the reaction proceed and wash the resin at the end, if it's a solid-phase reaction." However, he concedes that many researchers are currently using robotic systems instead of closed systems for combinatorial synthesis, "so the jury is out on which is the most acceptable."

CombiChem's instrument will be capable of automating both solid-phase and solution chemistry. "If you really want to exploit as much diversity as you can ... you have to be able to do something in addition to just solid-phase chemistry," says Myers. "The reason is pretty obvious. There are about 150 reactions

now that work on solid phase. Some of those reactions work extremely well, some are still very poor yielding. But the organic chemists have an armory of thousands of chemical reactions that have been developed over the years, and of course primarily most of those are done in solution."

3-Dimensional Pharmaceuticals Inc., Exton, Pa., is also developing combinatorial chemistry instrumentation. The system is based on a technique called DirectedDiversity, an iterative optimization process that explores combinatorial space through successive rounds of selection, combinatorial synthesis, and testing. In each step, a chemical library is generated by robotic instruments, structure-activity information is obtained on library members, and data are analyzed to determine how closely the synthesized compounds match a set of desired properties. In each succeeding iteration, the structure-activity models are refined and new compounds are created until desired drug leads have been identified.

Future needs and prospects

Combinatorial chemistry has come a long way in the past few years, but many challenges still lie ahead. For example, Ellman foresees further development of solid-support chemistry, including new linkage strategies and novel methods for synthesizing support-bound libraries and cleaving compounds from supports. "And people will continue to focus on different types of templates - novel templates for the versatile display of functionality," he says.

Ellman also believes "there are some interesting opportunities in the area of combining combinatorial strategies with computational strategies and structure-based design. The idea is to use information about three-dimensional structures of receptors and enzymes in combination with libraries to rapidly identify high-affinity ligands. It is going to be interesting to see how best to combine these two approaches." The recent SH3 study by Schreiber's group exemplifies this strategy of using a knowledge of protein and protein-ligand structure to help design optimal libraries.

Eric M. Gordon, vice president of research and director of chemistry at Affymax, points out that "some people enter into the molecular diversity sphere with the idea that it's a random process and that what you want to do is make as many molecules as you can that are as different from each other as they can be, but that no particular thought has to go into the design of these libraries. That's an extreme position. I believe that combinatorial chemistry doesn't stand alone. It should be integrated in with the arsenal of tools used for drug discovery, as opposed to being viewed as a competitive technology, say with structure-based design. I think the fate of it and the greatest power of it is going to be when it's used in concert with structure-based approaches and computational approaches."

Still believes that future prospects for combinatorial chemistry are good. "A lot of drugs will be discovered with it," he says. "And it's hard to imagine that there will not be people who find some enormously interesting catalytic compounds and stoichiometric reagents using these methods." However, he says, "The real key to making it work is twofold. First, you really need to have a good idea - a good basic structure that has a real chance of doing something really interesting, and you want to manufacture that idea in as many variants as you can afford to screen."

A second and even greater challenge, says Still, is devising novel and effective assays. "You need assays for the property you want that can be run in parallel, so you can select the beads or the library members that you want just by simple inspection. You simply look at them and pull out the ones that have the right property."

Gordon agrees that greater assay development is needed. "The amount of molecular diversity and the number of molecules that are going to become available are going to dwarf the present screening

capacities," he says. "What's required is more individualized assays and assay miniaturization."

Miniaturization not only saves on the cost of proteins and other rare materials used in assays but also provides greater compatibility with the very small amounts of combinatorial compounds that are typically synthesized on beads. "Instead of 96-well microtiter dishes, which are the current standard in the pharmaceutical industry, you're going to see 1,000-well trays," Gordon predicts.

Ontogen's Moran believes that an increased focus on analytical chemistry is needed in the combinatorial field. "One needs to produce a reasonable amount of material in order to characterize any one compound that might be of interest in a library, either by mass spectroscopy or more preferably by proton NMR," says Moran. "The reason for this is that organic synthesis is not straightforward."

Different groups added to a core structure will affect the reactivity of library members to a differential extent, leading to possible failures of key synthetic steps. "So one needs to get a handle on how much of each component is being produced and whether you're actually making all the components in your library," he says. "Unless we have good analytical control over our experiments, it's going to be a challenge knowing what one's made." However, another researcher comments that, although it's important to be able to analyze hits, it's not practical or even desirable to analyze all library compounds.

Janda believes another key goal for the future is "to create a global library or universal library, where if you screen the library you'd find a hit for any type of target. It might not be a very potent hit, but you'd find a lead. People are trying to create a small library that would be very diverse that would give you leads to almost anything in the drug area. Some people call this combinatorial chemistry's Holy Grail."

However, Still says the universal library may prove to be as permanently elusive as the Holy Grail. In combinatorial chemistry, he says, "medicinal chemists ... design the first library of 1,000 to 100,000 compounds that has a good chance of acting on the target they are going after. Then they screen and make a new sublibrary based on the structure-activity relationships they find. I don't think any medicinal chemist would believe any single library of 10,000 compounds, no matter how carefully chosen, will contain leads for every medical target."

Schreiber suggests that combinatorial molecular recognition could become a fundamental tool for understanding protein function. One of the ultimate goals of the Human Genome Project is to discover the functions of human proteins, he says, and up to now this has been done with molecular biology techniques.

"Virtually all studies of the functions of proteins today involve making mutations in the genes that encode proteins and studying the effects," he explains. "This genetic approach to studying protein function is very powerful, but it is very slow and very inefficient. It's going to take centuries to study the function of all the proteins encoded by the human genome this way, and that's simply unacceptable."

In principle, this problem could be solved, he says, by using a "chemical genetics approach - where instead of making mutations in the gene encoding the protein you attack the protein itself by using organic ligands that bind to it." And such ligands can best be identified with combinatorial methods.

Hence, says Schreiber, "Chemical genetics could be the way in the future to solve the problem of protein function. There's a big advantage if you do it that way - because the very act of understanding protein function gives you a molecule that actually alters function. In terms of medical applications of the knowledge we seek, that's what one is ultimately trying to do."

Combinatorial chemistry, coupled to structural biology and cell biology, "is the most likely avenue to solve the protein binding problem," he says. "If we can combine those techniques, the consequences will be very exciting. It will lead to an era where biology is intimately coupled to chemistry, and where one might even say that chemistry, rather than genetics, will drive biology."

The ultimate usefulness of combinatorial chemistry for drug discovery and other applications remains to be proved. But Lilly's Kaldor - whose group developed by combinatorial means the CNS agent that has advanced to the clinic - is one researcher who is cautiously optimistic. "These techniques are more broadly applicable than crystal-structure-guided design methods because you don't have to have any knowledge of your receptor in order to apply them. ... You can develop a pharmacophore hypothesis much more quickly than you might have otherwise been able to do so. To date, we have used combinatorial chemistry for lead generation or lead optimization in over 50% of current Lilly projects and anticipate this percentage will increase with time."

Lilly's development of the CNS compound took less than two years from target identification to the beginning of clinical trials. This is "very fast," says Kaldor, "and we, of course, are being challenged by our management to repeat this success in every project we work on. ... It's a stunning example of what can be done if ... you apply combinatorial chemistry."

Return to Article Index

[ACS Home Page] [ACS Publications Division Page]

Predictive Carcinogenicity: A Model for Aromatic Compounds, with Nitrogen-Containing Substituents, Based on Molecular Descriptors Using an Artificial Neural Network

Giuseppina Gini* and Marco Lorenzini

Dipartimento di Elettronica e Informazione, Politecnico di Milano, I-20133 Milano, Italy

Emilio Benfenati, Paola Grasso, and Maurizio Bruschi

Department of Environmental Health Sciences, Istituto di Ricerche Farmacologiche "Mario Negri", I-20157 Milano, Italy

Received April 25, 1999

A back-propagation neural network to predict the carcinogenicity of aromatic nitrogen compounds was developed. The inputs were molecular descriptors of different types: electrostatic, topological, quantum-chemical, physicochemical, etc. For the output the index TD50 as introduced by Gold and colleagues was used, giving a continuous numerical parameter expressing carcinogenicity. From the tens of descriptors calculated, principal component analysis enabled us to restrict the number of parameters to be used for the artificial neural network (ANN). We used 104 molecules for the study. An $R_{cv}^2 = 0.69$ was obtained. After removal of 12 outliers, a new ANN gave an R_{cv}^2 of 0.82.

1. INTRODUCTION

Man is exposed to many chemicals of natural and synthetic origin. An urgent question concerns their potential negative effects on human health. To identify chemicals inducing toxicity and to limit the incidence of human cancers and other diseases, rodent bioassays are the principal methods used today. However, this approach is not altogether problem-free, on several accounts: (1) the cost of the assay (>1 million U.S. dollars per chemical); (2) the time needed for the tests (3−5 years); (3) ethical considerations and public pressure to reduce or eliminate the use of animals in research and testing;[1] (4) difficulties in the extrapolation to man.

We were interested in the prediction of carcinogenicity, but cancer is not a single disease. Several mechanisms are involved in the various processes leading to the different tumors. This makes the task of assessing the computational prediction particularly challenging. Dedicated expert systems have been employed for computerized prediction of carcinogenicity.[2,3] However, these have limitations.[1−3] These expert systems work mainly on the assumption that toxicity is linked to the presence of toxic residues, either defined by human experts or found by the expert system. In some cases, the expert systems also use some simple physicochemical parameters. A very recent book describes the state-of-the-art of the research in the prediction of toxicity.[4]

Another widespread approach for predicting toxicity relies on molecular descriptors, which refer to global properties or characteristics of the molecule. In recent years a huge increase in the number of studies of theoretical molecular descriptors has appeared in the literature, including their use in toxicity prediction.[5] In the case of expert systems chemical data can be handled in several formats, but with artificial neural network (ANN) molecular descriptors are more suitable, and indeed they have been used in the prediction of carcinogenicity with contrasting results.[6−8] In this study we consider the use of molecular descriptors as input to ANN for the prediction of carcinogenicity of aromatic compounds with nitrogen-containing substituents.

2. METHODS

2.1. Input and Output of the Model. In many cases the carcinogenicity of a compound is classified by activity. A numerical, continuous approach was introduced by Gold and colleagues.[9] Gold's database contains standardized results for carcinogenicity for more than 1200 chemicals; for each substance it reports the carcinogenicity on rat and mouse, expressed using the parameter TD50, which is the chronic dose rate that would give half the animals tumors within some standard experimental time—the "standard lifespan" for the species. The huge amount of information in the database and the quantitative homogeneous evaluation are two important advantages. This database was therefore adopted as the basis for selecting the output parameter for the neural network. In the present study, for each chemical we chose the lowest (i.e. most potent) TD50. For the purpose of homogeneity all data refer to the mouse.

We limited the chemical compounds to be evaluated to those containing an aromatic ring and a nitrogen linked to the aromatic ring, because our previous experience with a commercial expert system showed that several of the compounds classified incorrectly belonged to this category.[10] The category includes several chemical classes, such as nitrosamines, amides, amines, and nitro derivatives, etc. The list of 104 selected compounds, with their toxic activity, is given in Table 1.

For the output we transformed the TD50 as follows: output = log(MW × 1000/TD50) (MW = molecular

10.1021/ci9903096 CCC: $18.00
Published on Web 10/07/1999

Table 1. Chemical Names, CAS Number, and Experimental and Calculated Toxic Values of 104 Compounds

name	CAS no.	expt	pred	name	CAS no.	expt	pred
(*N*-6)-(methylnitroso)adenine		0.6665	0.4923	5-nitroacenaphthene	602-87-9	0.6194	0.6508
(*N*-6)-methyladenine	443-72-1	0.0000	0.4462	acetaminophen	103-90-2	0.0000	0.3215
1,5-naphthalenediamine	2243-62-1	0.5838	0.5713	AF-2	3688-53-7	0.5922	0.5664
1-(1-naphthyl)-2-thiourea	86-88-4	0.8274	0.6979	aniline·HCl	142-04-1	0.2679	0.2523
1-amino-2-methylanthraquinone	82-28-0	0.5516	0.6981	anthranilic acid	118-92-3	0.1737	0.1693
1-[(5-nitrofurfurylidene)amino]hydantoin	67-20-9	0.4588	0.4831	atrazine	1912-24-9	0.6881	0.6902
2,2′,5,5′-tetrachlorobenzidine	15721-02-5	0.5963	0.6738	azobenzene	103-33-3	0.7571	0.7360
2,2,2-trifluoro-*N*-[4-(5-nitro-2-furyl)-2-thiazolyl]acetamide	42011-48-3	0.7321	0.6992	benzidine·2HCl	531-85-1	0.7086	0.6738
				c.i. disperse yellow 3	2832-40-8	0.4769	0.4644
2,4,5-trimethylaniline	137-17-7	0.7129	0.6384	chloramben	133-90-4	0.3477	0.2602
2,4,6-trimethylaniline·HCl	6334-11-8	0.6498	0.6310	chlorambucil	305-03-3	1.0000	0.9094
2,4-diaminoanisole sulfate	39156-41-7	0.4965	0.4567	cinnamyl anthranilate	87-29-6	0.4017	0.4539
2,4-diaminotoluene·2HCl	636-23-7	0.5643	0.5146	d & c red no. 9	5160-02-1	0.4336	0.4384
2,4-dimethoxyaniline·HCl	54150-69-5	0.4257	0.4197	dacarbazine	4342-03-4	0.8653	0.5674
2,4-dinitrophenol	51-28-5	0.0000	−0.0145	dapsone	80-08-0	0.0000	0.4293
2,4-dinitrotoluene	121-14-2	0.0000	0.3873	fd & c red no. 4	4548-53-2	0.2512	0.2209
2,4-xylidine·HCl	21436-96-4	0.6608	0.5765	fd & c yellow no. 6	2783-94-0	0.2717	0.2126
2,5-xylidine·HCl	51786-53-9	0.4458	0.5227	fluometuron	2164-17-2	0.5344	0.4913
2,6-dichloro-*p*-phenylenediamine	609-20-1	0.4405	0.4430	formic acid 2-[4-(5-nitro-2-furyl)-2-thiazolyl]hydrazide	3570-75-0	0.7277	0.6196
2-(acetylamino)fluorene	53-96-3	0.7563	0.7638				
2-amino-4-(5-nitro-2-furyl)thiazole	38514-71-5	0.7243	0.6966	furosemide	54-31-9	0.4876	0.5560
2-amino-4-(*p*-nitrophenyl)thiazole	2104-09-8	0.7133	0.6690	hydrochlorothiazide	58-93-5	0.4514	0.5654
2-amino-4-nitrophenol	99-57-0	0.4384	0.4929	*m*-cresidine	102-50-1	0.5100	0.5057
2-amino-5-nitrophenol	121-88-0	0.3238	0.3026	*m*-phenylenediamine·2HCl	541-69-5	0.4844	0.4144
2-amino-5-nitrothiazole	121-66-4	0.0000	−0.0862	*m*-toluidine·HCl	638-03-9	0.3831	0.3642
2-aminoanthraquinone	117-79-3	0.4630	0.6501	melamine	108-78-1	0.3532	0.4286
2-aminodiphenylene oxide	3693-22-9	0.7344	0.7324	melphalan	148-82-3	0.9803	1.0032
2-biphenylamine·HCl	2185-92-4	0.4241	0.3075	methotrexate	59-05-2	0.6443	0.4927
2-chloro-*p*-phenylenediamine sulfate	61702-44-1	0.4001	0.4022	metronidazole	443-48-1	0.4927	0.4924
2-hydrazino-4-(5-nitro-2-furyl)thiazole	26049-68-3	0.6857	0.6391	mexacarbate	315-18-4	0.8264	0.8305
2-hydrazino-4-(*p*-aminophenyl)thiazole	26049-71-8	0.7018	0.6003	*N*-(1-naphthyl)ethylenediamine·2HCl	1465-25-4	0.0000	0.2226
2-hydrazino-4-(*p*-nitrophenyl)thiazole	26049-70-7	0.7134	0.6021	*N*-nitrosodiphenylamine	86-30-6	0.4952	0.4837
2-methyl-1-nitroanthraquinone	129-15-7	0.8404	0.7969	*N*-phenyl-*p*-phenylenediamine·HCl	2198-59-6	0.0000	0.4836
2-naphthylamine	91-59-8	0.6557	0.6456	*N*-[4-(5-nitro-2-furyl)-2-thiazolyl]-formamide	24554-26-5	0.7325	0.7051
2-nitro-*p*-phenylenediamine	5307-14-2	0.4532	0.2208				
2-*sec*-butyl-4,6-dinitrophenol	88-85-7	0.8360	0.8256	*N*-[5-(5-nitro-2-furyl)-1,3,4-thiadiazol-2-yl]acetamide	2578-75-8	0.7440	0.5990
3,3′-dimethoxybenzidine-4,4′-diisocyanate	91-93-0	0.2791	0.4109				
3-(3,4-dichlorophenyl)-1,1-dimethylurea	330-54-1	0.4788	0.6364	nithiazide	139-94-6	0.4609	0.4735
3-chloro-*p*-toluidine	95-74-9	0.3807	0.3849	nitrofen	1836-75-5	0.6198	0.5780
3-nitro-*p*-acetophenetide	1777-84-0	0.3995	0.4186	*o*-aminoazotoluene	97-56-3	0.5936	0.5913
4′-fluoro-4-aminodiphenyl	324-93-6	0.8306	0.5675	*o*-anisidine·HCl	134-29-2	0.4162	0.4160
4,4′-methylenebis(2-chloroaniline)·2HCl	64049-29-2	0.6141	0.6300	*o*-phenylenediamine·2HCl	615-28-1	0.4333	0.4379
4,4′-methylenebis(*N*,*N*-dimethyl)benzenamine	101-61-1	0.5456	0.5662	*o*-toluidine·HCl	636-21-5	0.4296	0.4531
4,4′-methylenedianiline·2HCl	13552-44-8	0.6760	0.6068	*p*-anisidine·HCl	20265-97-8	0.0000	0.3522
4,4′-oxydianiline	101-80-4	0.6680	0.5299	*p*-chloroaniline	106-47-8	0.3917	0.3951
4-amino-2-nitrophenol	119-34-6	0.0000	0.2578	*p*-cresidine	120-71-8	0.5986	0.5234
4-aminodiphenyl	92-67-1	0.8312	0.6604	*p*-isopropoxydiphenylamine	101-73-5	0.4703	0.4558
4-chloro-*m*-phenylenediamine	5131-60-2	0.4088	0.3924	*p*-nitrosodiphenylamine	156-10-5	0.5024	0.6000
4-chloro-*o*-phenylenediamine	95-83-0	0.4233	0.4286	*p*-phenylenediamine·2HCl	624-18-0	0.3813	0.3249
4-chloro-*o*-toluidine·HCl	3165-93-3	0.6942	0.6084	pentachloronitrobenzene	82-68-8	0.6161	0.6816
4-nitro-*o*-phenylenediamine	99-56-9	0.0000	0.3094	phenacetin	62-44-2	0.2859	0.3255
4-nitroanthranilic acid	619-17-0	0.2882	0.3812	phenylhydrazine	100-63-0	0.0000	0.0428
5-nitro-2-furaldehyde semicarbazone	59-87-0	0.6600	0.4923	proflavine·HCl hemihydrate	952-23-8	0.6535	0.6667
5-nitro-*o*-anisidine	99-59-2	0.4276	0.4530	pyrimethamine	58-14-0	0.5199	0.6243

weight), in order to have a more continuous output space and to refer to the moles of the chemical, not the weight.[8]

2.2. Molecular Descriptors. Chemical structures were drawn with Hyperchem (Hypercube, Inc.) and optimized using the PM3 Hamiltonian. We used the following programs to calculate descriptors: VAMP version 6.1 (Oxford Molecular Ltd.) for the quantum-mechanical and thermodynamic calculations, on a Silicon Graphics XS24 workstation; HAZARD EXPERT version 3.0 (CompuDrug Chemistry Ltd., Budapest, Hungary) for log *D* calculation; TSAR version 3.0 (Oxford Molecular) for the other descriptors, using a personal computer.

We calculated the 34 descriptors listed in Table 2. log *D* was calculated at pH 2, 7.4, and 10 as representative of the pH of the stomach, blood, and gut, where different processes may occur with the chemicals. The complete set of values is available from the authors on request.

Table 2. The 34 Used Descriptors

molecular weight	three principal axes of inertia
log *D* at pH 2, 7.4, 10	Balaban Index
HOMO	Wiener Index
LUMO	Randic Index
heat of formation	five Kier & Hall connectivity indices
dipole moment	
polarizability	six Kier shape indices
total energy	flexibility index
molecular volume	ellipsoidal volume
three principal moments of inertia	electrotopological sum

2.3. Reducing the Number of Descriptors by Principal Component Analysis. We used principal component analysis

(PCA) to select a smaller set of descriptors so the network could converge faster.

The main change in the set of 104 molecules (accounting for 63% of the total variability) was explained by the descriptor total energy and by a pool of descriptors including topological, geometric, and electrostatic values inversely correlated with the first principal component (PC). The second PC, accounting for another 8% of the variability, was mainly related to the dipole moment, the topological index of Balaban, and the quantum-chemical HOMO and LUMO descriptors and to log D at pH 7.4 and pH 10. The log D at pH 2 correlated with the third PC, thus explaining another smaller but different source of variability.

Descriptors with the highest scores on the first four components of PCA (accounting for 85% of the total variability) were chosen and reduced, eliminating those most closely correlated. A final criterion was to keep a pool of descriptors representing the different aspects of the molecule considered (physicochemical, electronic, and topological, etc). From the 34 descriptors calculated, 13 were selected: molecular weight, HOMO, LUMO, dipole moment, polarizability, Balaban, ChiV3 and flexibility indices, log D at pH 2 and pH 10, third principal axis of inertia, ellipsoidal volume, and electrotopological sum.

2.4. Artificial Neural Networks. In all the simulations, performed with MBP v 1.1,[11] the working parameters were set as follows: the weight initialized with the SCAWI technique; net gain $\eta(0) = 0.75$; initial moment $\alpha(0) = 0.9$; acceleration factor YPROP, $K_a = 0.7$, $K_d = 0.07$. The algorithm stopped itself when it encountered one of the following conditions: gradient lower than 10^{-6}; mean square error in validation (MSE) equal to 0; maximum calculated difference between calculated and desired output equal to 0; maximum number of iterations reached. Each network was trained starting from 100 random points in space, in order to minimize the probability of converging toward local minima. Input data were scaled between 0 and 1 in order to have a homogeneous range of variation of descriptors. The output was scaled accordingly.

For the validation step the leave-two-out approach was adopted, i.e. a cross-validation procedure using two examples in validation and the others for training. Five ANN models were generated, using data sets composed of 84 molecules randomly chosen in the training set and 20 in the test set.

The software is available on request, for noncommercial use.

3. RESULTS AND DISCUSSION

Most QSAR studies consider a limited number of parameters, taking account of previous knowledge in the field and using multivariate linear analysis. In our case there was no previous knowledge on the importance of specific molecular descriptors. We therefore considered a wide range of different classes, as detailed above, to extract information without a priori elimination of any possibilities.

We tried using regression analysis, but without success. ANN can be used to model complex phenomena where noise and nonlinear processes may be present, such as in our case. A disadvantage is the time needed, because many iterations are needed. This is a weakness of this neural network if we want to keep all the descriptors as inputs. Reducing the inputs shortens the training time. If this involves eliminating redundancy, the net has more chance of finding relevant parameters. For these reasons (reduction of computation time and elimination of redundancy) we chose PCA to select the ANN inputs, as it has been used for this purpose in other cases.[12,13] A risk related to the use of PCA is the possibility of eliminating inputs which behave nonlinearly. To verify that we had not eliminated any useful information, we built a new ANN using the first 12 PCs as inputs. These contain about 99% of the information of the original set of variables. The results with these PCs were comparable to those with the selected descriptors, shown below, indicating that we had not lost information through our selection.

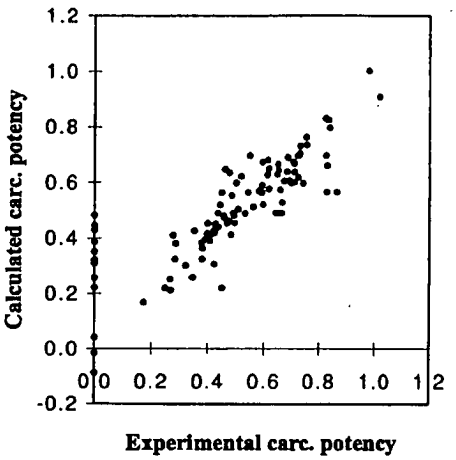


Figure 1. Predicted versus experimental carcinogenicity values with the BPNN four-neuron model.

Table 3. Results with BPNN of Increasing Numbers of Hidden Neurons (MSE = Mean Square Error)[a]

neurons	MSE	R_{cv}^2	neurons	MSE	R_{cv}^2
3	0.0157	0.675	6	0.0153	0.676
4	*0.0146*	*0.691*	7	*0.0146*	*0.691*
5	0.0154	0.676			

[a] Best results are in italics.

Table 1 gives results of the back-propagation neural network (BPNN). Figure 1 shows the predicted and experimental carcinogenicity with the BPNN four-neuron model.

The average R^2 cross-validated (R_{cv}^2) after 10 000 iterations and using different numbers of internal neurons is shown in Table 3.

To overcome possible representation bias in our data set, we built up five random data sets composed of 84 molecules in the training set and 20 in the test set. Then five independent ANN models were generated. This approach has been used recently.[14] R_{cv}^2 for these models was 0.70 using four or six neurons in the inner layer, in agreement with the leave-two-out method (see Table 3).

For the BPNN, the presence of outliers in the set was assumed and investigated in order to see whether the network's capacity for generalization improved after removing them and to assess the chemical nature of the activity of the compounds.

We adopted a conservative approach to remove outliers, taking only the molecules presenting an error in validation higher than 0.2 in the two best models (those with four and

Table 4. Results with BPNN of Increasing Number of Hidden Neurons, after Removing the Outliers (MSE = Mean Square Error)[a]

neurons	MSE	R_{cv}^2	neurons	MSE	R_{cv}^2
3	0.0062	0.793	6	0.0057	0.810
4	*0.0053*	*0.824*	7	0.0061	0.792
5	0.0053	0.824	8	0.0073	0.755

[a] Best results are in italics.

seven internal neurons). Twelve molecules were identified as outliers and removed. The results are presented in Table 4.

The results show that R_{cv}^2 has been clearly improved. Most of the outliers (9 out of 12) are molecules for which the experimental results for carcinogenicity were not statistically significant and an arbitrary value of 10^{31} was given in the Gold database (see Figure 1; they lie on the y axis, because of the transformation formula described in section 2.1 and scaling). The main experimental evidence for these molecules suggests noncarcinogenicity. Other considerations on the outliers regard their homogeneity from a chemical point of view. As we said, the compounds used for this ANN belong to several chemical classes and the outliers appear to be distributed over various chemical classes. Some are chemicals that have no structures in common with other members of the set, and this may explain their behavior. However, the ANN correctly predicted the toxicity of other chemicals which appear badly represented.

Special consideration must be given to two molecules, *o*- and *p*-anisidine. These isomers have identical or very similar chemical descriptors. However, their toxicity is very different, due to different metabolism in the animals. The ANN based on molecular descriptors was not able to distinguish them. This is a case of interesting behavior, shared with other compounds, which may undergo a metabolic process able to detoxify the chemical. In another study we solved the case of *o*- and *p*-anisidine by an expert system which distinguishes the toxic substructure.[15]

The present study illustrates the possibilities and limitations of the approach based on molecular descriptors. From the chemical point of view *o*- and *p*-anisidine may appear very similar, but for a living organism they are not. There are, however, chemicals which appear different within various chemical classes—as in the case of the compounds we have used—that the organism considers similar, because they are converted to aromatic amines. Knowledge of the body's bioprocesses is therefore an important source of information. Knowledge of the structural features of the molecule that characterize its specific mechanism of action cannot be ignored in some cases, in order to solve problems occurring in the prediction.

Another general point is the reliability of the database. We used an authoritative database, resulting from critical assessment of data from two sources: reports in the literature using different experimental protocols and results obtained according to a uniform protocol within the U.S. National Toxicology Program. Differences in the sources may affect the homogeneity of the data.[16] Furthermore, this database, like many others, changes constantly as new studies appear, adding knowledge.

A final comment on the database is that in most cases it still contains a limited number of compounds (despite the huge amount of work needed to build them up), so for some compounds we did not have enough examples to train the ANN properly.

4. CONCLUSIONS

Many models for toxicity prediction use linear relationships, which apply well within congeneric chemical classes. ANN has been used in limited cases. Villemin et al. used ANN to model polycyclic aromatic compounds in carcinogenic classes, obtaining good results.[6] Vracko obtained an r of 0.83, after removing the outliers, for a set of aromatic compounds belonging to different chemical classes.[8] Benigni and Richard, in a study using 280 compounds of various kinds, concluded that BPNN models fitted the training sets but had no general applicability.[7] The main feature of their study is the large differences between the structures of the molecules, much wider than in the other ANN used to predict carcinogenicity, including our present study.

The present study shows the feasibility of an ANN for predicting carcinogenicity of chemicals of various types. Several chemical classes are in fact present.

Our study attempts to illustrate how knowledge can be improved using ANN, probably because it is modeling nonlinearity. With chemical descriptors as input ANN is useful for cases where multilinear regression fails. We are aware of the limitations of this approach, which are common to other methods, as discussed. However, we believe that no single approach can cope with the vast problem of predictive toxicology, as already noted by other authors.[17] The next task is the extension to a wider set of chemicals. How to extract rules from the ANN is a major topic, and how to integrate ANN results with those from independent sources. We have already evaluated this last point in some cases, coupling expert systems and ANN within hybrid systems able to incorporate the best elements from each of the approaches.[15,18]

ACKNOWLEDGMENT

We acknowledge the financial support of the European Commission (Grant ERB-CP94 1029 until 1998 and Grant ENV4-CT97-0508 since 1998) and NATO (Grant CRG 971505), from 1998. We thank Dr. Y.-t. Woo for useful discussion.

REFERENCES AND NOTES

(1) Omenn, G. S. Assessing the Risk Assessment Paradigm. *Toxicology* **1995**, *102*, 23−28.
(2) Benfenati, E.; Gini, G. Computational Predictive Programs (Expert Systems) in Toxicology. *Toxicology* **1997**, *119*, 213−225.
(3) Dearden, J. C.; Barratt, M. D.; Benigni, R.; Bristol, D. W.; Combes, R. D.; Cronin, M. T. D.; Judson, P. N.; Payne, M. P.; Richard, A. M.; Tichy, M.; Worth, A. P.; Yourick, J. J. The Development and Validation of Expert Systems for Predicting Toxicity. *ATLA* **1997**, *25*, 223−252.
(4) Gini, G. C.; Katritzky, A. R. *Predictive Toxicology of Chemicals: Experiences and Impact of AI Tools*; AAAI Press: Menlo Park, CA, 1999; pp 1−155.
(5) Karelson, M.; Maran, U.; Wang, Y.; Katritzky, A. R. QSPR and QSAR Models Derived with CODESSA Multipurpose Statistical Analysis Software. In *Predictive Toxicology of Chemicals: Experiences and Impact of AI Tools; AAAI 1999 Spring Symposium Series*; Gini, G. C., Katritzky, A. R., Eds.; AAAI Press: Menlo Park, CA, 1999; pp 45−48.
(6) Villemin, D.; Cherqaoui, D.; Mesbash, A. Predicting Carcinogenicity of Polycyclic Aromatic Hydrocarbons from Back-Propagation Neural Network. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1288−1293.

(7) Benigni, R.; Richard, A. M. QSARS of Mutagens and Carcinogens: Two Case Studies Illustrating Problems in the Construction of Models for Noncongeneric Chemicals. *Mutat. Res.* **1996**, *371*, 29−46.

(8) Vracko, M. A Study of Structure−Carcinogenic Potency Relationship with Artificial Neural Networks. The Using of Descriptors Related to Geometrical and Electronic Structures. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1037−1043.

(9) Gold, L. S.; Slone, T. H.; Manley, N. B.; Backman Garfinkel, G.; Hudes, E. S.; Rohrbach, L.; Ames, B. N. The Carcinogenic Potency Database: Analyses of 4000 Chronic Animal Cancer Experiments Published in the General Literature and by the U.S. National Cancer Institute/National Toxicology Program. *Environ. Health Perspect.* **1991**, *96*, 11−15.

(10) Benfenati, E.; Tichy, M.; Malvè, L.; Grasso, P.; Gini, G. Expert Systems for Toxicity Prediction Based on Fragment Recognition: Evaluation of a Commercial System and Improved Approaches; American Chemical Society Meeting, Las Vegas, NV, Sep 8−12, 1997; Abstracts of paper, COMP 136.

(11) Anguita, D. *Matrix Back Propagation v 1.1: User's Manual*; 1993. Available through anonymous ftp at risc6000.dibe.unige.it.

(12) Miao, X.; Azimi-Sadjadi, M. R.; Tina, B.; Dubey, A. C.; Witherspoon, N. H. Detection of Mines and Minelike Targets Using Principal Component and Neural-Network Methods. *IEEE Trans. Neural Networks* **1998**, *9*, 454−463.

(13) Ventura, S.; Silva, M.; Pérez-Bendito, D.; Hervás, C. Computational Neural Networks in Conjunction with Principal Component Analysis for Resolving Highly Nonlinear Kinetics. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 287−291.

(14) Sussman, N. B.; Macina, O. T.; Claycamp, H. G.; Grant, S. G.; Rosenkranz, H. S. The Utility of Multiple Random Sampling in the Development of SAR Models. In *Predictive Toxicology of Chemicals: Experiences and Impact of AI Tools*; *AAAI 1999 Spring Symposium Series*; Gini, G. C., Katritzky, A. R., Eds.; AAAI Press: Menlo Park, CA, 1999; pp 45−48.

(15) Gini, G.; Lorenzini, M.; Vittore, A.; Benfenati, E.; Grasso, P. Some Results for the Prediction of Carcinogenicity Using Hybrid Systems. In *Predictive Toxicology of Chemicals: Experiences and Impact of AI Tools*; *AAAI 1999 Spring Symposium Series*; Gini, G. C., Katritzky, A. R., Eds.; AAAI Press: Menlo Park, CA, 1999; pp 139−143.

(16) Helma, C.; Gottmann, E.; Kramer, S.; Pfahringer, B. Data Quality Issues in Toxicological Knowledge Discovery. In *Predictive Toxicology of Chemicals: Experiences and Impact of AI Tools*; *AAAI 1999 Spring Symposium Series*; Gini, G. C., Katritzky, A. R., Eds.; AAAI Press: Menlo Park, CA, 1999; pp 8−11.

(17) Bahler, D.; Bristol, D. W. Prediction of Chemical Carcinogenicity in Rodents by Machine Learning of Decision Trees and Rule Sets. In *Predictive Toxicology of Chemicals: Experiences and Impact of AI Tools*; *AAAI 1999 Spring Symposium Series*; Gini, G. C., Katritzky, A. R., Eds.; AAAI Press: Menlo Park, CA, 1999; pp 74−77.

(18) Gini, G.; Testaguzza, V.; Benfenati, E.; Todeschini, R. HyTEx (Hybrid Toxicology Expert System): Architecture and Implementation of a Multi-domain Hybrid Expert System for Toxicology. *Chemom. Intell. Lab. Syst.* **1998**, *43*, 135−145.

CI9903096